Document anonymization for border guards and immigration services

Henri Bouma [1*], Raimon Pruim [1], Arthur van Rooijen [1], Johan-Martijn ten Hove [1], Jelle van Mil [1], Ben Kromhout [2]

[1] TNO Intelligent Imaging, The Hague, The Netherlands
[2] IND Document and Identity Investigation Office, Zwolle, The Netherlands

ABSTRACT

The current capabilities and capacities of border guards and immigration services can be enhanced using technologies that automate the analysis of travel, identity and breeder documents in order to detect fraud. These technologies can be relevant for countering emerging threats in document and identity verification (e.g., forged documents, impostor fraud, morphed faces) at both manual and highly automated border control points (both in the first and in the second line) and in the issuance process of genuine documents. The travel documents (e.g., passports) and breeder documents (e.g., birth certificates) contain personal information, such as name, date of birth and national number. The personal information must be well protected and a data breach must be avoided at all times. One of the ways to protect the personal data is to minimize the sharing of personal data. Anonymization removes the personal information (e.g., by replacing the personal information by a black bar) and can therefore be used to minimize the sharing of personal data. This paper describes the tool that assists border guards and immigration services for the anonymization of travel and breeder documents. The tool consists of a graphical user interface, document detection, keyword recognition, face detection, number detection, barcode detection and masking of personal data. The results show that only 10 annotated images are needed to reach a keyword detection accuracy of 96% and anonymization sensitivity of 93% of the related personal data.

Keywords: Anonymization, travel documents, breeder documents, border guards, immigration services

1. INTRODUCTION

The current capabilities and capacities of border guards and immigration services can be enhanced using technologies that automate the analysis of travel, identity and breeder documents in order to detect fraud. The technology is important to counter threats in identity and document verification (such as morphed faces, forged documents or impostor fraud) at border control points and in the document issuance process. The travel documents (e.g., passports) and breeder documents (e.g., birth certificates) contain personal information, such as name, date of birth and national number. The personal information must be well protected and a data breach must be avoided at all times. One of the ways to protect the personal data is to minimize the sharing of personal data (GDPR article 5-c). Anonymization removes the personal information (e.g., by replacing the personal information by a black bar) and can therefore be used to minimize the sharing of personal data (GDPR article 25-1). Development of modules for automatic processing of travel and breeder document requires examples of these documents. Therefore, it is necessary to share the documents with the developers and the documents contain personal data. However, sharing of personal data should be minimized. The minimization can be done in two ways.

The first way is that anonymization is applied before the document processing (Figure 1). Personal data is removed in an early stage and only anonymized documents are shared with developers. The effect is that border guards or immigration services can record data and anonymize the data before sharing it with researchers that develop modules for document processing. The advantage of this way is that even researchers cannot access the personal data, which is optimal for privacy protection. For some document processing modules the personal data may be irrelevant and it may be sufficient to see other details on the document.

* Henri.Bouma@tno.nl; phone +31 888 66 4054; www.tno.nl

H. Bouma, R. Pruim, A. van Rooijen, J.M. ten Hove, J. van Mil, B. Kromhout, "Document anonymization for border guards and immigration services," Proc. SPIE, vol. 11542, (2020). https://doi.org/10.1117/12.2571944

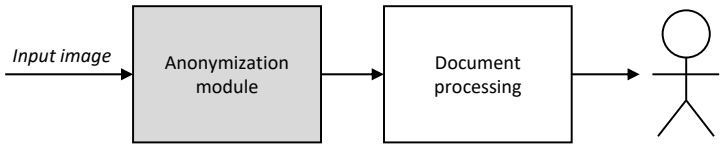

Figure 1: Use of anonymization before document processing.

The second way is that anonymization is applied after document processing (Figure 2). In some cases the processing requires raw input images that have not been modified by an anonymization module. For example, a document authentication module that verifies whether the names are manipulated must be able to inspect those regions. In this case it is necessary to share the document including the personal information to the researchers, and the researchers should take appropriate actions to protect the data. Although it is necessary to process the raw input images, it may not be necessary to show the complete image on the screen. For example, during field tests, demonstrations or dissemination activities, it may be sufficient to show anonymized images and thereby minimize the sharing of personal data to a restricted audience (e.g., evaluators or other end users).

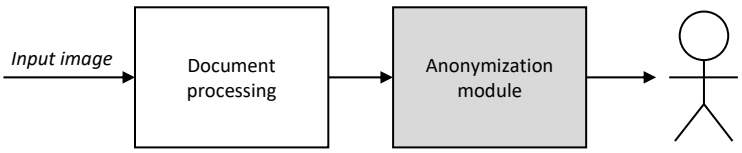

Figure 2: Use of anonymization after document processing.

This document describes the tool for automated anonymization of travel documents and breeder documents. Section 2 shows the user workflow. Section 3 describes the method. Section 4 presents the results. Finally, the conclusions are summarized in Section 5.

2. USER WORKFLOW

The anonymization tool is intended to be used for travel or breeder documents that contain personal data. To avoid undesired transfer of personal data from one organization to another organization, the anonymization tool is provided with an empty database and with untrained models. Typically, the anonymization tool needs several manually annotated examples of the same document type before it can automatically anonymize that type of document. On one hand new annotations and training are required to use the tool. On the other hand, the amount of annotations is minimized and only a few examples are needed (see Section 4 for results). The principal users will be personnel of border guard organizations and immigration services. The system has the capability to detect faces, barcodes, numbers and keywords in the scanned image of the document. This enables the system to localize personal information in the scanned image and replace the personal information by a black bar. The tool needs several manually annotated examples before it can anonymize automatically.

The interactive user procedure consists of the following steps:

1. Load data (one or multiple scans).
2. Verify country and doc-type, and optionally correct.
3. Verify document rotation, and optionally correct.
4. Verify the anonymization, and optionally correct.
5. Export the anonymized image and go to the next scan.

The graphical user interface is shown in Figure 3. Keywords are indicated in green and masks are indicated in red. The first document can be anonymized in random order. All subsequent documents should be anonymized in the same order, to recognize keywords and locations consistently. The ‘Id’ and the ‘Snippets’ in the center help to annotate in a consistent way, even when the user is not able to translate the snippet. For example, if in the first document the snippet with ID=1 is related to the name, then also in the second document, the snippet with ID=1 should correspond to the name. Snippets are extracted from the first document where they are created.

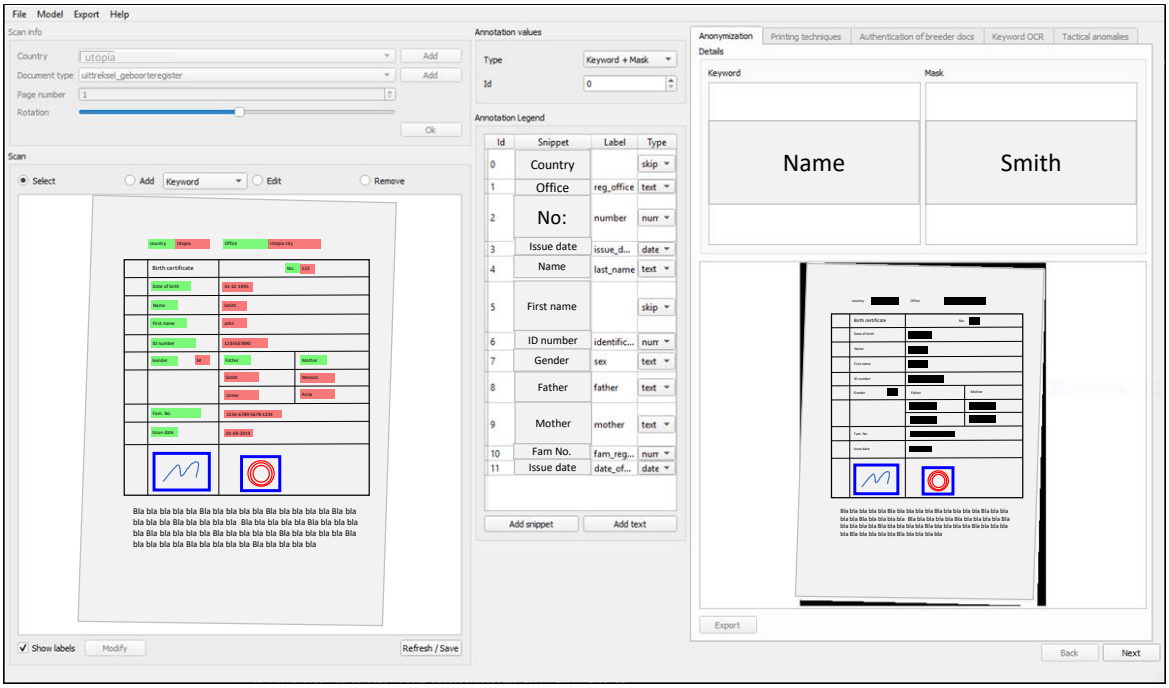

Figure 3: Graphical user interface.

3. METHOD

The documents may contain different types of personal data. The four most important categories are the following:

- Photograph: Facial information from a person.
- Barcode: barcode that may contain personal information that is easy to read for machines.
- Number without keyword: document number without any neighboring keyword.
- Pair of keyword and value: The document may contain pairs of keywords and values. For example:
    - Keyword = “Last name”, Value = “Smith”
    - Keyword = “Date of Birth”, Value = “31-12-2001”

Each of the categories is recognized in a different way by the anonymization module.

3.1 Internal architecture

The internal architectural overview is shown in Figure 4.

3.2 Document recognition

The module for “document recognition” automatically recognizes the country name and the document type (e.g., “identity card” or “birth certificate”). The module is implemented with VGG16 [Simonyan, 2015][*], which has been chosen because it is an easy-to-use open-source deep-learning algorithm for image classification.

3.3 Document rotation

The module for “document rotation” automatically detects the document orientation and uses this calculated angle to put the document upright. This module is implemented using the Hough Line Transform and Canny Edge detector from OpenCV[†]. The calculated document angle is the median angle of all the found horizontal lines (max +/- 15 degrees) in the document.

[*] https://keras.io/applications/
[†] https://opencv.org

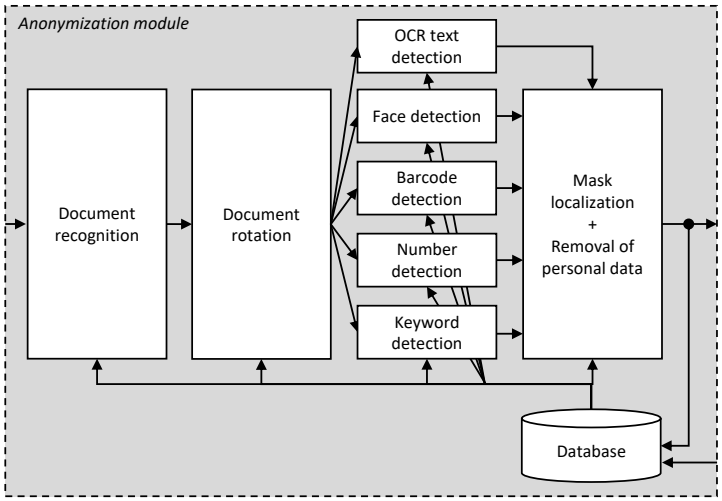

Figure 4: Internal architecture.

3.4 Face detection

The module for "face detection" automatically detects and localizes a photograph. The module is implemented with a specific DLIB face detector[*], which was chosen because it is a commonly-used state-of-the-art open-source algorithm for face analysis.

3.5 Barcode detection

The module for "barcode detection" automatically detects and localizes barcodes of different types (1D-barcodes or 2D-barcodes such as QR-code or PDF417). The module is implemented with a Faster R-CNN [Ren, 2017] , which is a state-of-the-art deep-learning algorithm for object detection and classification [Boer, 2017].

3.6 Number detection

The module for "number detection" can localize numbers even when there are no keywords in the neighborhood. The number detection is implemented with TESSERACT[†], which was chosen because it one of the most accurate open-source optical character recognition (OCR) engines. The module generates multiple strings (words) on a document. If the string contains multiple digits, then the string is recognized as a number.

3.7 Keyword detection

Keywords (such as 'First name' or 'Last name' or 'Date of birth') can be detected with a specific keyword detector – which detects the keyword as a whole. Keyword detection is implemented with a Faster R-CNN [Ren, 2017]. The pre-processing consists of data augmentation by copying the relevant keyword to other locations in the same document image and therefore generating multiple artificial document images. The augmentation uses color transformation and blending.

3.8 OCR text detection

Text detection is performed using the CRAFT text detector [Baek, 2019] as implemented in the open Keras-OCR engine[‡]. This implementation localizes words in the images. This detector was chosen instead of TESSERACT since the CRAFT detector is considered more suitable for detecting arbitrarily-oriented and curved text. The OCR text detection is used to improve the localization of masks.

3.9 Mask localization and removal of personal data

The personal information in an image must be masked to generate anonymized images. Masks for the face, barcode and number are trivial to implement, because the location of the mask is identical to the positioning information generated by

[*] http://dlib.net
[†] https://github.com/tesseract-ocr/
[‡] https://github.com/faustomorales/keras-ocr

the (face, barcode or number) detection. However, the pairs of keyword and value are less trivial, because the location of the value (that contains personal information) should be derived from the detected keywords. The method is implemented in two steps.

The first step calculates the distance between every mask-keyword combination in every document image in the set of verified document images. Because the document images are of varying resolutions, the distances used in this process are relative to the page height and width of the document. Some languages read from left to right and other languages read from right-to-left. In this distance calculation, the top-left coordinate (x1, y1) of the keyword is used as anchor point if the keyword is on the right side of the mask (reading right-to-left), and the bottom-right coordinate (x2, y2) otherwise. When all relative distances are calculated, the 5th percentile (top-left coordinate of the mask) and 95th percentile (bottom-right coordinate of the mask) are used as a 'template offset' for that mask-keyword combination.

The second step calculates the location of the masks for a new document, using the detected keywords, the template offsets and OCR text-detections:

1. For every keyword that should be in the document, it is checked if the keyword is detected. If the keyword is detected at the expected place, the direct keyword-mask offset is used. If a keyword is not detected, or if the detected keyword is not near the expected place in the document, the keyword-mask offset from the nearest keyword is used.
2. The bounding box of a mask is calculated by first retransforming the relative offset to an absolute offset and then adding the offsets to the anchor point of the detected keyword.
3. The resulting bounding box of the mask is optimized using OCR text-detections to fit the mask more closely around the text. Only the OCR text detections are valid that (1) do overlap with the mask and (2) do not overlap with the keyword and (3) with a larger width than height. The outer coordinates of these valid OCR detections are used to optimize the location of the anonymization mask. The resulting anonymization mask may become larger (e.g., if the OCR detections are partially overlapping) or smaller (e.g., if the OCR detections are inside the mask).

Finally, the pixels in the image that are located inside the masks are replaced by black pixels, thereby removing the personal information from the image.

4. RESULTS

This section gives an overview of the experimental setup and results. The experiments focus on technical accuracy of the components for document recognition, face detection, barcode detection, keyword recognition, number detection and masking.

4.1 Experimental setup

The anonymization tool is tested on a limited set of data. Five different document types are used and each document type has 20 scans (of 20 different documents), so the total dataset contains 100 images. In each image, faces, barcodes, numbers, keywords and personal information are marked with boxes. The used documents contain different alphabets (including Latin and Arabic) and different levels of standardization: highly standardized (identity card) and less standardized (birth/civil register).

4.2 Document recognition

Document recognition was trained on a large database with many document classes. The performance of document recognition is validated on the dataset with 100 images. The results are shown as a confusion matrix in Table 1, where ground-truth is indicated as 'gt' and system predictions are indicated as 'pred'. The average document recognition accuracy is 92%.

Table 1: Confusion matrix for document recognition.

	Country A (pred)	Country B (pred)	Country C (pred)	Country D (pred)	Country E (pred)
Country A (gt)	20	0	0	0	0
Country B (gt)	0	20	0	0	0
Country C (gt)	0	8	12	0	0
Country D (gt)	0	0	0	20	0
Country E (gt)	0	0	0	0	20

4.3 Face detection

The performance of the pretrained face detector is validated on the 40 faces in the dataset. The results are shown in Table 2, where FPR indicates the false-positive rate. The table shows that the system detects all faces without any false positives.

Table 2: Results for face detection.

Number of Faces	Accuracy	FPR
40	100%	0%

4.4 Barcode detection

The barcode detector is trained on a part of the dataset and tested on another part of the dataset. This dataset contains 307 different files with 449 unique barcodes from 11 country and document type combinations. The barcode types are 1D, QR and PDF417. This dataset is very challenging because it also contains barcodes printed on the back side of the documents. This is possible since the documents have ink on both sides. Also these barcodes can easily be missed by a human observer as they are very faint. It is still important to detect and anonymize them, since they do contain sensitive information and a dedicated adversary could still extract data from them.

The detector is trained on 276 files containing 408 barcodes and evaluated on the remaining 31 documents with 41 barcodes. A barcode is considered to be detected (True Positive) if the intersection-over-union (IoU) between the detection and annotation is higher than 0.7.

The results are shown in Table 3, where TP is true positive, FP is false positive and FN is false negative. There are 41 barcodes, of which 40 are correctly detected (TP) and only 1 barcode is missed (FN). Furthermore, there is one false detection (FP) at a location where there is actually not a barcode present. The results show that almost 98% of the barcodes is correctly detected. We want to stress that the FP and FN are a consequence of wanting to have a network which is able to detect faint barcodes printed on the backside of the documents.

Table 3: Results for barcode detection.

Number of Barcodes (TP+FN)	TP	FP	FN
41	40	1	1

4.5 Keyword detection

The keyword detector is trained on a part of the dataset and tested on another part of the dataset. The detector is evaluated on documents from three countries (A, B, C). A keyword is considered to be detected (True Positive) if the intersection-over-union (IoU) between the detection and annotation is higher than 0.5. The Average Precision (AP) of the detections is defined as accuracy. This detector is evaluated with leave-N-document-out cross-validation. The results are shown in Table 4. The results show that only 5 annotated images are needed to reach a keyword detection accuracy of 92% and 10 images are needed to reach a keyword detection accuracy of 96%.

Table 4: Results for keyword detection.

No. training samples per doc. type	Keyword detection accuracy (AP)
1	75%
3	84%
5	92%
10	96%

4.6 Number detection

The number detector is applied to detect the document numbers in the identity cards that contain a Non-Latin document number without a related keyword. The results are shown in Table 5. The table shows that approximately 45% of these numbers are detected (TP) and 55% are missed (FN). The single false positive (FP) was actually not a number but part of an icon. Note that detected numbers related to a keyword (e.g., birth date) are ignored in the statistics (i.e., are considered neither as a FP or TP). The 45% is not very high, but may be acceptable for end-users during interactive annotation, because we can infer the number mask from other keyword locations.

Table 5: Results for non-Latin number detection.

Number of Numbers (TP+FN)	TP	FP	FN
9	4	1	5

4.7 Masking textual data

For the evaluation of the masking it is not sufficient to solely define whether the data field has been 'detected' or not. Rather, it needs to be assessed to which extent the personal data areas have been covered, and the extent of falsely anonymized areas. Therefore we define two measures for accuracy assessment of the masking:

1) True Positive Rate (i.e. sensitivity or recall): TPR = TP / (TP + FN)
2) False Positive Rate: FPR = FP / (TP + FP),

where the correct anonymization is indicated as true positive area (TP), the incorrectly missed area is indicated as false negative (FN), and the incorrectly anonymized area is indicated as false positive (FP). The results are shown in Table 6. These results are obtained using the keyword detector over two train/test splits of the data for each document type (i.e., per document type train on 10 annotated images, test on the 10 remaining images, and vice versa). The table shows that the average anonymization sensitivity (TPR) is 96% when the masks are related to ground-truth keyword boxes and the average anonymization sensitivity is 93% when the masks are related to automatic keyword boxes, which means that almost all masks are placed correctly. The average FPR is 35%, which indicates some over-segmentation. For anonymization, it is better to have over-segmentation than to reveal the personal information, and therefore the trade-off between FPR and TPR is chosen in such that the FPR is further from 0 (over segmentation) than the TPR is from 100 (missing a region with personal information). Figure 5 shows a representative example with an FPR of 35% to give an impression of the amount of over-segmentation, where the yellow boxes are the manual masks and the black boxes are the automatic masks (photo and number are white and keywords are blue).

5. CONCLUSIONS

Document anonymization is important to minimize the sharing of personal data. The tool described in this paper can be used for anonymization of travel and breeder documents. The tool consists of a graphical user interface, document recognition, keyword detection, face detection, number detection, barcode detection and masking of personal data. The results show that only 10 annotated images are needed to reach a keyword detection accuracy of 96% and an anonymization sensitivity of 93% of the related personal data. Face and barcode detection also reached a high accuracy of 100% and 98% respectively. Only the OCR-based number detection has lower performance but can be compensated by approximate localization of masks based on keywords. Based on the tolerance for errors, which is typically "none", there should always be a manual inspection of these results.

Table 6: Results for masking.

	Masking with ground-truth keyword boxes		Masking with automatic keyword detection	
	TPR	**FPR**	**TPR**	**FPR**
Country A	99%	34%	98%	31%
Country B	94%	31%	82%	40%
Country C	92%	44%	91%	43%
Country D	98%	33%	98%	33%
Country F	98%	34%	97%	34%
TOTAL	**96%**	**35%**	**93%**	**36%**

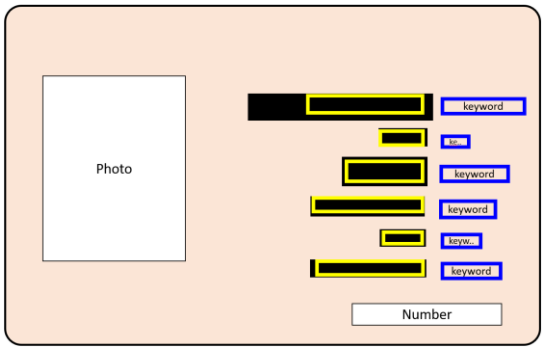

Figure 5: Example result of masking textual data with FPR=35%.

ACKNOWLEDGEMENT

The work described in this paper is performed in the H2020 project D4FLY ("Detecting Document frauD and iDentity on the fly"). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 833704.

REFERENCES

[1] Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. "Character Region Awareness for Text Detection," IEEE CVPR, (2019).
[2] Boer M. de, Bouma, H., Kruithof, M., et al., "Automatic analysis of online image data for law enforcement agencies by concept detection and instance search," Proc. SPIE 10441, (2017).
[3] Hermans, A., Beyer, L., & Leibe, B., "In defense of the triplet loss for person re-identification," arXiv:1703.07737, (2017).
[4] Ren, S., He, K., Girshick, R., Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", IEEE Trans. Pattern Analysis and Machine Intelligence 39, 1137-1149 (2017).
[5] Simonyan, K., & Zisserman, A., "Very deep convolutional networks for large-scale image recognition," ICLR, (2015).